

***i*-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study**

Kunlin Zhang, Sijia Cui, Suhua Chang, Liuyan Zhang and Jing Wang*

Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 100101, Beijing, China

Received February 12, 2010; Revised April 3, 2010; Accepted April 17, 2010

ABSTRACT

Genome-wide association study (GWAS) is nowadays widely used to identify genes involved in human complex disease. The standard GWAS analysis examines SNPs/genes independently and identifies only a number of the most significant SNPs. It ignores the combined effect of weaker SNPs/genes, which leads to difficulties to explore biological function and mechanism from a systems point of view. Although gene set enrichment analysis (GSEA) has been introduced to GWAS to overcome these limitations by identifying the correlation between pathways/gene sets and traits, the heavy dependence on genotype data, which is not easily available for most published GWAS investigations, has led to limited application of it. In order to perform GSEA on a simple list of GWAS SNP *P*-values, we implemented GSEA by using SNP label permutation. We further improved GSEA (*i*-GSEA) by focusing on pathways/gene sets with high proportion of significant genes. To provide researchers an open platform to analyze GWAS data, we developed the *i*-GSEA4GWAS (*improved* GSEA for GWAS) web server. *i*-GSEA4GWAS implements the *i*-GSEA approach and aims to provide new insights in complex disease studies. *i*-GSEA4GWAS is freely available at <http://gsea4gwas.psych.ac.cn/>.

INTRODUCTION

Genome-wide association study (GWAS) has become a popular approach by utilizing genome-wide genotyping

array to map susceptibility effects through examining associations between SNP genotype frequency and traits (1,2). The standard GWAS analysis focuses on single SNP/gene and identifies only a number of the most significant SNPs that account for only a small proportion of the genetic variants and offers limited understanding of complex diseases (3). It also ignores the combined SNP/gene effects representing interactions of multiple genetic factors of complex disease, which lead to difficulties in exploring biological function and mechanism from a systems point of view.

To overcome these limitations, the principle of gene set analysis (GSA), which was recently developed to facilitate handling of genome-wide expression data (4), has been introduced to GWAS. GSA tests the correlation between gene set, which is defined as genes involved in the same pathway, and trait, resting on the assumption that genes belonging to a disease-related pathway/gene set tend to reveal concordant signals. Recently, several implementations or application programs using the GSA principle have been developed (3,5–14). An interesting implementation is the gene set enrichment analysis (GSEA) which technically evaluates whether the distribution of genes sharing a biochemical or cellular function is different from the distribution of a ranked genome-wide gene list by computing Kolmogorov–Smirnov like statistics (3,15). Successful applications of GSEA in GWAS include the study of Parkinson's disease (3), Crohn's disease (11) and the study of host control for HIV-1 (9).

However, GSEA usually depends on genotype data, which is not easily available for most published GWAS investigations. This leads to a limited application of it, especially application in the secondary analysis of published GWASs to best interpret the remaining long list of GWAS data from a systems point of view. In order to perform GSEA on easily available GWAS data,

*To whom correspondence should be addressed. Tel: 86 10 6485 5841; Fax: 86 10 6485 5841; Email: wangjing@psych.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

mainly SNP P -values, we implemented GSEA by using SNP label permutation instead of phenotype label permutation to analyze P -values. We further improved the GSEA (*i*-GSEA) by focusing on pathways/gene sets with high proportions of significant genes instead of relying only on the total significance coming from either a few or many significant genes. Our study shows that *i*-GSEA has the improved sensitivity to identify pathways/gene sets representing combined effects of possibly modest SNPs/genes.

There have been more than 480 GWAS publications covering a broad range of human diseases available by the end of January 2010 [see <http://www.genome.gov/gwastudies/> and (16)] and the number keeps increasing. In order to implement the *i*-GSEA approach and provide researchers with an open platform to analyze the easily available GWAS data from both publications and self experiments, we developed the *i*-GSEA4GWAS (*improved* GSEA for GWAS) web server. *i*-GSEA4GWAS aims to identify novel pathways/gene sets correlated with given traits and provide new insights into complex disease study, especially in complementation with standard single SNP/gene based GWAS analysis.

Overview of the *i*-GSEA algorithm

The *i*-GSEA4GWAS web server implements *i*-GSEA to help researchers explore GWAS data efficiently. *i*-GSEA is an application and extension of GSEA. For GSEA, it runs the following three key procedures. (i) The max statistics or $-\log(P\text{-value})$ of closely spaced SNPs in a gene is used to represent the gene; then, the ranked gene list with corresponding representing values is utilized to calculate each gene set's enrichment score (ES), i.e. a Kolmogorov–Smirnov like statistics with weight 1, which reflects the trend that genes of a gene set tend to be located at the top of the entire ranked genome-wide gene list. (ii) The phenotype label permutation (to break the association between genotype and phenotype) and a straightforward normalization are performed to generate the distribution of the ES and correct gene variation (i.e. different genes with different number of SNPs mapped will result in identification of gene sets containing genes with more SNPs mapped, instead of genes with functional correlation) and gene set variation (i.e. different gene sets contain different number of genes) (3). (iii) Based on all the distributions of ES s generated by permutation, false discovery rate (FDR) is used for multiple testing correction (17).

In *i*-GSEA, we implement SNP label permutation instead of phenotype label permutation to analyze SNP P -values and to correct gene and gene set variation and multiply k/K to the ES to get the significance proportion based enrichment score ($SPES$), where k is the proportion of significant genes of the gene set and K is the proportion of significant genes of the total genes in the GWAS. Here, significant genes are defined as the genes mapped with at least one of the top 5% of all SNPs. Instead of ES which focuses on the total significance coming from either a few or many significant genes, $SPES$ emphasizes on total significance coming from high proportion of significant genes. So, *i*-GSEA trends to pick

up pathways/gene sets including a high proportion of significant genes and is more appropriate for study of the combined effects of possibly modest SNPs/genes in complex disease. This ensures *i*-GSEA the improved sensitivity.

Description of the *i*-GSAE4GWAS web server

System development, configuration and system overview. *i*-GSEA4GWAS is written in Java and JSP. The program is running in a high-end computing cluster with six dedicated Intel XEON Quad-Core CPUs at 2.0 GHz and two dedicated Intel XEON Quad-Core CPUs at 3.0 GHz (representing a total of 32 cores) with a total of 96 GB of memory.

i-GSEA4GWAS is freely available and there is NO registration required. We also provide a template program for users to access our web server through command lines instead of web browser. Figure 1 shows the system architecture and overview of our *i*-GSEA4GWAS web server.

Input data. The input data of *i*-GWAS are SNP data, mainly P -values. The server will transform the P -value to $-\log(P\text{-value})$. Some other values of association test like odds ratio and statistics are also accepted. Here, we define all these input values as 'association value'. The detailed format of input is a SNP list with two columns, namely SNP identifier and its corresponding association value. A gene list with gene HUGO symbol (<http://www.genenames.org/>) (18) and along with corresponding association value is also accepted but not recommended due to the limitation of gene variation as described above.

The collection of pathways/gene sets utilized in *i*-GSEA4GWAS are extracted and curated from MSigDB v2.5 (<http://www.broadinstitute.org/gsea/msigdb/>) (15) to ensure comprehensiveness and high quality. It includes gene sets denoting canonical pathways integrated from a variety of online resources like KEGG (19) and BioCarta (<http://www.biocarta.com/>), and GO (gene ontology) terms with high confidence (20). Additionally, *i*-GSEA4GWAS supports customized gene sets, i.e. users can upload their own gene sets. This provides the opportunity that the knowledge database (gene sets) can always be expanded, and makes it easy for users to test their hypothesis within their research interest.

Running procedure. First, our server maps SNPs to the genome-wide genes. *i*-GSEA4GWAS contains an optional multiple-level broad-to-narrow SNPs->genes mapping system, which helps users either fully explore genome-wide SNP data (by choosing '500-kb upstream and downstream of gene') or narrow down the research scope by only focusing on functional SNPs (by choosing 'functional SNP'). The SNPs->genes mapping is established based on SNP and gene annotations from the Ensembl BioMart database (Release 56–15 September 2009, <http://www.ensembl.org/biomart/martview>) (21). Next, the program uses the list of SNP-mapped genes to filter the collection of pathways/gene sets to obtain candidate pathways/gene sets. Additional filtering includes a

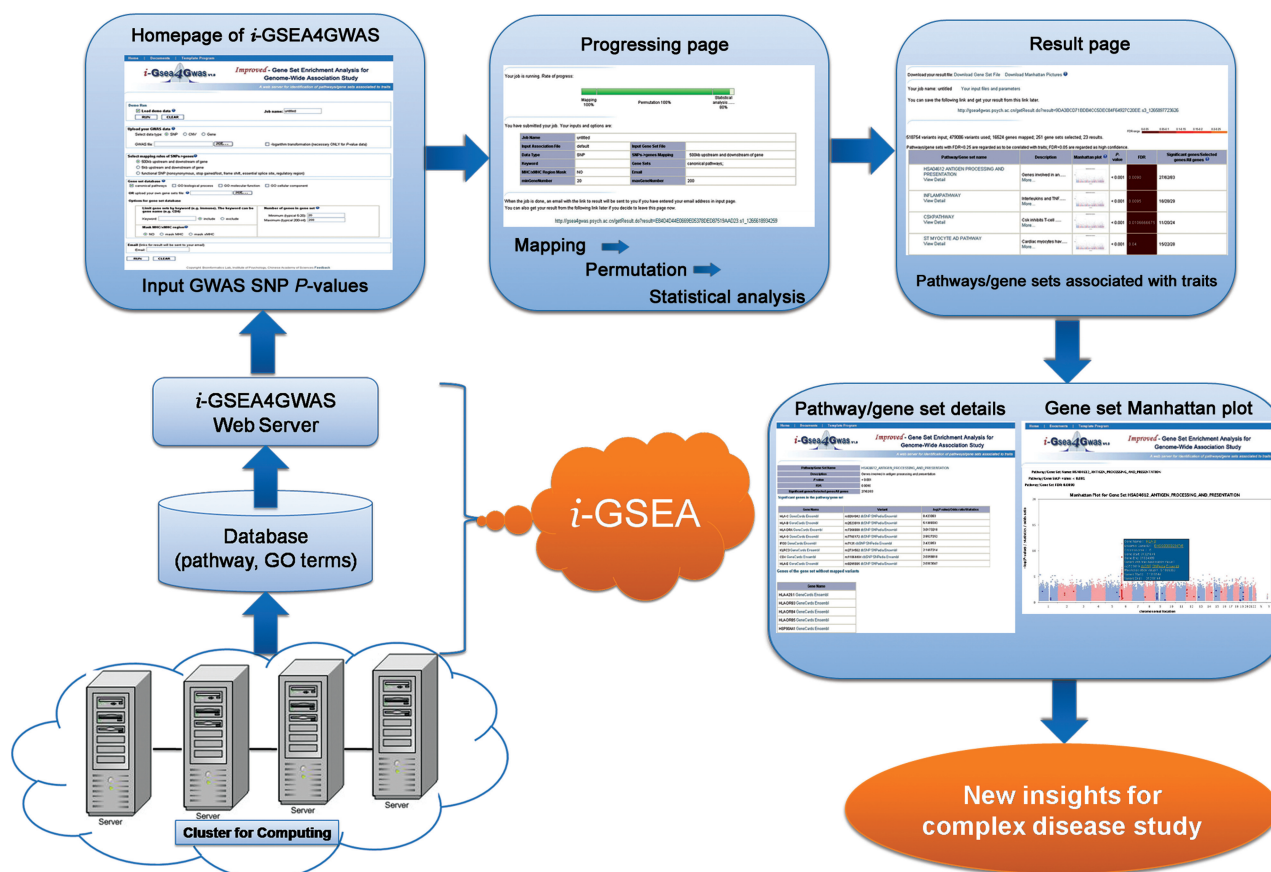


Figure 1. The system architecture and overview of *i-GSEA4GWAS* web server.

keyword search to define the function of gene sets, to mask major histocompatibility complex/extended major histocompatibility complex (MHC/xMHC) region (22) to only explore the non-MHC/xMHC region (23), and to restrain gene set size to avoid overly narrow or overly broad functional categories and too large a search space, with the default parameter of minimum 20 and maximum 200 (3,9). Then *i-GSEA4GWAS* performs *i-GSEA* to identify pathways/gene sets correlated with traits. When the job is running, the progress of the three key steps (mapping, permutation and statistical analysis) will be shown in the progress bar.

Output and examples. The pathways/gene sets, with $FDR < 0.25$ (the threshold of $FDR < 0.25$ denotes the confidence of ‘possible’ or ‘hypothesis’, while the threshold of $FDR < 0.05$ is regarded as ‘high confidence’ or ‘with statistical significance’), are displayed in a tabular form with links to detailed text information of pathways/gene sets, genes and SNPs. One characteristic feature of *i-GSEA4GWAS* output is the Manhattan plot of gene set, which uses the Manhattan plot of GWAS as background, and highlights the results of association test for a given pathway/gene set. It helps users to compare the association results of the given pathway/gene set with the genome-scale data graphically. The result page also contains the download URL, from where all the results, both texts and figures, can be downloaded.

The first example data is SNP *P*-value from the GWAS of HIV-1 host control investigation, which studied the quantitative trait—HIV-1 viral load set point and used Illumina HumanHap550 BeadChip (24). The input includes 518 754 SNPs with corresponding *P*-values. There are 479 086 SNPs mapped to 16 524 genes by using the default ‘500 kb upstream and downstream of gene’. 251 gene sets were selected by using ‘canonical pathways’, and using the default gene set size ranges from 20 through 200. A total of 23 pathways/gene sets were identified to be possibly correlated with the trait ($FDR < 0.25$). Among the 23 pathways, five pathways namely ‘antigen processing and presentation’, ‘inflammation’, ‘cskpathway’, ‘ribosome’ and ‘st myocyte ad pathway’ have high confidence ($FDR < 0.05$). The pathways ‘antigen processing and presentation’ ($FDR = 0.009$), ‘inflammation’ ($FDR = 0.010$), and ‘cskpathway’ ($FDR = 0.011$) have been replicated in our follow-up work of the initial GWAS by either the top hits or using GSEA (9). Both the pathways ‘ribosome’ and ‘st myocyte ad pathway’ were only identified by *i-GSEA* ($FDR < 0.05$) and were not replicated by the follow-up GWAS and GSEA analysis. However, both pathways have references to support their relevance to HIV. For pathway ‘ribosome’ ($FDR = 0.047$), it has been known that the ribosomal frame shifting is a kind of gene expression mechanism of

several RNA viruses (including HIV-1 and SARS-CoV) to express replicase enzymes and a couple of investigations have been carried out on this topic (25–27). The pathway ‘st myocyte ad pathway’ (FDR = 0.040) contains a variety of adrenergic receptors that induce subtype-specific signaling effects. The top gene with the smallest *P*-value in this gene set, *ITPR1* (inositol 1,4,5-triphosphate receptor, type 1), has been reported to interact with genes in HIV. For example HIV-1 *Nef* interacts with *ITPR1* to trigger the activation of plasma membrane calcium influx channels (28); HIV-1 *Tat* induces release of calcium from *ITPR1*-regulated stores in neurons and astrocytes, an effect that plays an important role in Tat-induced TNF- α production (29).

The second example data are SNP *P*-values of the GWAS of bipolar disorder, which is a case/control study and utilized the GeneChip 500K Mapping Array Set (Affymetrix chip) (2). *i*-GSEA4GWAS identified two pathways/gene sets (FDR < 0.05) by using the default ‘500 kb upstream and downstream of gene’ SNPs->genes mapping rule, searching ‘canonical pathways’ and GO terms, and masking the xMHC region. One pathway is ‘glutathione metabolism’ (FDR = 0.036), which was also reported by Holmans *et al* (30). Reducing glutathione (GSH) level will lead to an increased potential for cellular oxidative stress, which has been implicated in the pathology of bipolar disorder, and recently some studies have been performed on animal models (31,32). The other one is ‘cysteine type endopeptidase activity’ (FDR = 0.041), which might be a novel functional biological mechanism related to bipolar disorder and worthy of further study. There is no pathway/gene set found with FDR < 0.05 or FDR < 0.25 by using GSEA with the same parameters.

DISCUSSION

Our web server implements *i*-GSEA for GWAS data analysis to explore the biological function and mechanism from a systems point of view. It is efficient not only for in-house GWASs with available genotype data, but also for secondary data mining of published GWASs, where genotype data are not readily available. *i*-GSEA4GWAS aims to help researchers take the best advantage of the high-cost GWAS data to look for new biological insights worthy of further confirmation. This represents the major potential for *i*-GSEA4GWAS. The improved *i*-GSEA approach, the optimal multiple-level SNPs->genes mapping system, the comprehensive pathway/gene set database with options of customization, as well as the consideration of MHC/xMHC region distinguish *i*-GSEA4GWAS from other programs. It should be noted that *i*-GSEA4GWAS does not take into account the linkage disequilibrium (LD) patterns from SNP arrays and would not prune the set of SNPs for LD since these can only be done when genotype data is available. So, users are recommended to input SNPs not in LD (say $r^2 < 0.2$) to reduce the possibility of biased results due to LD patterns from SNP arrays. On the

other hand, as the name of the web server defines, *i*-GSEA4GWAS is only applicable to whole-genome SNP arrays.

i-GSEA employs SNP label permutation to correct gene variation to reduce the bias due to different genes with different number of mapped SNPs, which is ignored by the non-randomization approach such as the segmentation test (33) as implemented in the GeSBAP web server (10). This correction ensures to identify gene sets consisting of non-random high-association genes with biological plausibility instead of random high-association genes with large numbers of mapped SNPs. Furthermore, the key issue of *i*-GSEA, focusing on pathways/gene sets with high proportion of significant genes to detect combinations of modest effects, greatly improves sensitivity. Let n denote the number of pathways with FDR < 0.05 and m represent the number of pathways with both FDR < 0.05 and references to support, our comparison study between *i*-GSEA and GSEA shows, in the HIV-1 host control and bipolar disorder study as described above, $n = 7$ and $m = 6$ for *i*-GSEA, while $n = 2$ and $m = 2$ for GSEA. Further comparison by using the other six Wellcome Trust Case Control Consortium GWASs except bipolar disorder (2) obtained the result that $n = 33$ and $m = 14$ for *i*-GSEA, while $n = 2$, $m = 1$ for GSEA. In addition, the pathways identified (FDR < 0.05) by *i*-GSEA include all the pathways identified by GSEA in all the above comparisons. These show that *i*-GSEA has improved sensitivity in comparison to GSEA. Our data examples also show that *i*-GSEA4GWAS obtains very meaningful results with different GWAS designs (quantitative trait and case/control) and different genotyping platforms (Illumina and Affymetrix).

There are always conflicts between the amount and the quality of the gene sets utilized for computation. Using a large amount of gene sets, such as all GO terms, will introduce large background and noise due to the large number and low confidence of some gene sets, while only using a few gene sets will lose information. Moreover, the knowledge of gene sets is keeping increasing. To overcome this, we chose a balanced strategy, using a curated pathway/gene set database with consideration of both comprehensiveness and high quality, and allow researchers to upload their customized gene sets to make sure the gene sets of research focus will be well represented. This ensures that *i*-GSEA4GWAS contains a high-quality and reasonable search space. The *i*-GSEA4GWAS will be regularly updated to ensure the most up-to-date searching database and annotations. *i*-GSEA4GWAS also integrates the curated copy number variations (CNVs) from the Database of Genomic Variants (<http://projects.tcag.ca/variation/>) (34). In the future, the CNV functional module will be further extended to include the CNV probes of Illumina, Affymetrix and more genome-wide genotyping arrays.

In summary, the *i*-GSEA4GWAS web server provides researchers an efficient open platform for GWAS analysis, helping further interpret the SNP *P*-values from hundreds of available GWASs and future GWASs to provide new insights into disease study.

ACKNOWLEDGEMENTS

We thank Dr Lei Kong and Dr Ge Gao from Peking University for their kind help on the system configurations. We thank all our colleagues and friends in Chinese Academy of Sciences, who helped test the web server and provided us valuable suggestions. We thank the anonymous reviewers for their helpful comments and suggestions.

FUNDING

Project for Young Scientists Fund, Institute of Psychology, Chinese Academy of Sciences (O9CX115011); National Natural Science Foundation of China (NSFC) (30700441); and the Beijing New Star Project, Beijing Municipal Science & Technology Commission Foundation (2007A082). Funding for open access charge: O9CX115011.

Conflict of interest statement. None declared.

REFERENCES

- McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P. and Hirschhorn,J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
- Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
- Wang,K., Li,M. and Bucan,M. (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
- Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
- Holden,M., Deng,S., Wojnowski,L. and Kulle,B. (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**, 2784–2785.
- Torkamani,A., Topol,E.J. and Schork,N.J. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, **92**, 265–272.
- Askland,K., Read,C. and Moore,J. (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.*, **125**, 63–79.
- Chen,L., Zhang,L., Zhao,Y., Xu,L., Shang,Y., Wang,Q., Li,W., Wang,H. and Li,X. (2009) Prioritizing risk pathways: a novel association approach to searching for disease pathways using SNPs and pathways. *Bioinformatics*, **25**, 237–242.
- Fellay,J., Ge,D., Shianna,K.V., Colombo,S., Ledergerber,B., Cirulli,E.T., Urban,T.J., Zhang,K., Gumbs,C.E., Smith,J.P. *et al.* (2009) Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.*, **5**, e1000791.
- Medina,I., Montaner,D., Bonifaci,N., Pujana,M.A., Carbonell,J., Tarraga,J., Al-Shahrour,F. and Dopazo,J. (2009) Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res.*, **37**, W340–W344.
- Wang,K., Zhang,H., Kugathasan,S., Annese,V., Bradfield,J.P., Russell,R.K., Sleiman,P.M., Imielinski,M., Glessner,J., Hou,C. *et al.* (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am. J. Hum. Genet.*, **84**, 399–405.
- Yu,K., Li,Q., Bergen,A.W., Pfeiffer,R.M., Rosenberg,P.S., Caporaso,N., Kraft,P. and Chatterjee,N. (2009) Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.*, **33**, 700–709.
- Peng,G., Luo,L., Siu,H., Zhu,Y., Hu,P., Hong,S., Zhao,J., Zhou,X., Reveille,J.D., Jin,L. *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.*, **18**, 111–117.
- Cantor,R.M., Lange,K. and Sinsheimer,J.S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **86**, 6–22.
- Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
- Hindorf,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
- Reiner,A., Yekutieli,D. and Benjamini,Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.
- Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.*, **34**, D319–D321.
- Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
- Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
- Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
- Horton,R., Wilming,L., Rand,V., Lovering,R.C., Bruford,E.A., Khodiyar,V.K., Lush,M.J., Povey,S., Talbot,C.C. Jr, Wright,M.W. *et al.* (2004) Gene map of the extended human MHC. *Nat. Rev. Genet.*, **5**, 889–899.
- Warrens,A. and Lechler,R. (1999) *HLA in Health and Disease*. Academic Press, San Diego, CA.
- Fellay,J., Shianna,K.V., Ge,D., Colombo,S., Ledergerber,B., Weale,M., Zhang,K., Gumbs,C., Castagna,A., Cossarizza,A. *et al.* (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science*, **317**, 944–947.
- Mazauric,M.H., Seol,Y., Yoshizawa,S., Visscher,K. and Fourmy,D. (2009) Interaction of the HIV-1 frameshift signal with the ribosome. *Nucleic Acids Res.*, **37**, 7654–7664.
- Gendron,K., Charbonneau,J., Dulude,D., Heveker,N., Ferbeyre,G. and Brakier-Gingras,L. (2008) The presence of the TAR RNA structure alters the programmed -1 ribosomal frameshift efficiency of the human immunodeficiency virus type 1 (HIV-1) by modifying the rate of translation initiation. *Nucleic Acids Res.*, **36**, 30–40.
- Brierley,I. and Dos Ramos,F.J. (2006) Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res.*, **119**, 29–42.
- Manninen,A. and Saksela,K. (2002) HIV-1 Nef interacts with inositol trisphosphate receptor to activate calcium signaling in T cells. *J. Exp. Med.*, **195**, 1023–1032.
- Mayne,M., Holden,C.P., Nath,A. and Geiger,J.D. (2000) Release of calcium from inositol 1,4,5-trisphosphate receptor-regulated stores by HIV-1 Tat regulates TNF-alpha production in human macrophages. *J. Immunol.*, **164**, 6538–6542.
- Holmans,P., Green,E.K., Pahwa,J.S., Ferreira,M.A., Purcell,S.M., Sklar,P., Owen,M.J., O'Donovan,M.C. and Craddock,N. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.
- Dean,O.M., van den Buuse,M., Bush,A.I., Copolov,D.L., Ng,F., Dodd,S. and Berk,M. (2009) A role for glutathione in the pathophysiology of bipolar disorder and schizophrenia? Animal models and relevance to clinical practice. *Curr. Med. Chem.*, **16**, 2965–2976.

32. Dean,O., Bush,A.I., Berk,M., Copolov,D.L. and van den Buuse,M. (2009) Glutathione depletion in the brain disrupts short-term spatial memory in the Y-maze in rats and mice. *Behav. Brain Res.*, **198**, 258–262.
33. Al-Shahrour,F., Arbiza,L., Dopazo,H., Huerta-Cepas,J., Minguéz,P., Montaner,D. and Dopazo,J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.
34. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.